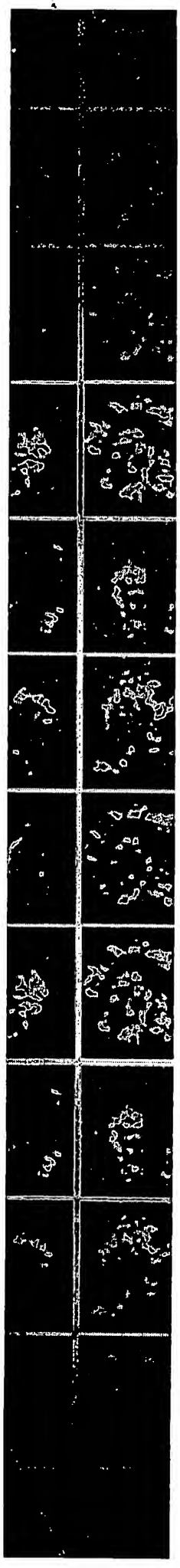



# Mass spectrometry-based proteomics

Ruedi Aebersold* & Matthias Mann†

*Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904, USA (e-mail: raebersold@systemsbiology.org)
†Center for Experimental BioInformatics (CEBI), Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark (e-mail: mann@bmb.sdu.dk)

**Recent successes illustrate the role of mass spectrometry-based proteomics as an indispensable tool for molecular and cellular biology and for the emerging field of systems biology. These include the study of protein–protein interactions via affinity-based isolations on a small and proteome-wide scale, the mapping of numerous organelles, the concurrent description of the malaria parasite genome and proteome, and the generation of quantitative protein profiles from diverse species. The ability of mass spectrometry to identify and, increasingly, to precisely quantify thousands of proteins from complex samples can be expected to impact broadly on biology and medicine.**

Proteomics in general deals with the large-scale determination of gene and cellular function directly at the protein level. But as the accompanying articles in this issue describe, the field is a collection of various technical disciplines, all of which contribute to proteomics. These include cell imaging by light and electron microscopy, array and chip experiments, and genetic readout experiments, as exemplified by the yeast two-hybrid assay. Another powerful proteomic approach focuses on the *de novo* analysis of proteins or protein populations isolated from cells or tissues. Such studies typically pose challenges owing to the high degree of complexity of cellular proteomes and the low abundance of many of the proteins, which necessitates highly sensitive analytical techniques. Mass spectrometry (MS) has increasingly become the method of choice for analysis of complex protein samples. MS-based proteomics is a discipline made possible by the availability of gene and genome sequence databases and technical and conceptual advances in many areas, most notably the discovery and development of protein ionization methods, as recognized by the 2002 Nobel prize in chemistry.

Here we survey the state of the field, particularly as it has evolved over the three years since the last review in these pages[1]. Already, many of the dreams of the discipline have at least been partly realized. MS-based proteomics has established itself as an indispensable technology to interpret the information encoded in genomes. So far, protein analysis (primary sequence, post-translational modifications (PTMs) or protein–protein interactions) by MS has been most successful when applied to small sets of proteins isolated in specific functional contexts. The systematic analysis of the much larger number of proteins expressed in a cell, an explicit goal of proteomics, is now also rapidly advancing, due mainly to the development of new experimental approaches.

Today, proteomics still remains a multifaceted, rapidly developing and open-ended endeavour. Although it has enjoyed tremendous recent success, proteomics still faces significant technical challenges. Each breakthrough that either allows a new type of measurement or improves the quality of data made by traditional types of measurements expands the range of potential applications of MS to molecular and cellular biology. Indeed, this field is already too expansive for a comprehensive, single review; thus we apologize in advance for the many omissions. However, we do hope that this article captures the excitement of recent achievements in MS-based proteomics, and points the way towards the direction future developments will likely take.

## Principles and instrumentation

Mass spectrometric measurements are carried out in the gas phase on ionized analytes. By definition, a mass spectrometer consists of an ion source, a mass analyser that measures the mass-to-charge ratio ($m/z$) of the ionized analytes, and a detector that registers the number of ions at each $m/z$ value. Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are the two techniques most commonly used to volatize and ionize the proteins or peptides for mass spectrometric analysis[2,3]. ESI ionizes the analytes out of a solution and is therefore readily coupled to liquid-based (for example, chromatographic and electrophoretic) separation tools (Fig. 1). MALDI sublimates and ionizes the samples out of a dry, crystalline matrix via laser pulses. MALDI-MS is normally used to analyse relatively simple peptide mixtures, whereas integrated liquid-chromatography ESI-MS systems (LC-MS) are preferred for the analysis of complex samples.

The mass analyser is, literally and figuratively, central to the technology. In the context of proteomics, its key parameters are sensitivity, resolution, mass accuracy and the ability to generate information-rich ion mass spectra from peptide fragments (tandem mass or MS/MS spectra) (see Fig. 1 and refs 1,4,5). There are four basic types of mass analyser currently used in proteomics research. These are the ion trap, time-of-flight (TOF), quadrupole and Fourier transform ion cyclotron (FT-MS) analysers. They are very different in design and performance, each with its own strength and weakness. These analysers can be stand alone or, in some cases, put together in tandem to take advantage of the strengths of each (Fig. 2).

In ion-trap analysers, the ions are first captured or 'trapped' for a certain time interval and are then subjected to MS or MS/MS analysis. Ion traps are robust, sensitive and relatively inexpensive, and so have produced much of the proteomics data reported in the literature. A disadvantage of ion traps is their relatively low mass accuracy, due in part to the limited number of ions that can be accumulated at their point-like centre before space-charging distorts their distribution and thus the accuracy of the mass measurement. The 'linear' or 'two-dimensional ion trap'[6,7] is an exciting recent development where ions are stored in a cylindrical volume that is considerably larger than that of

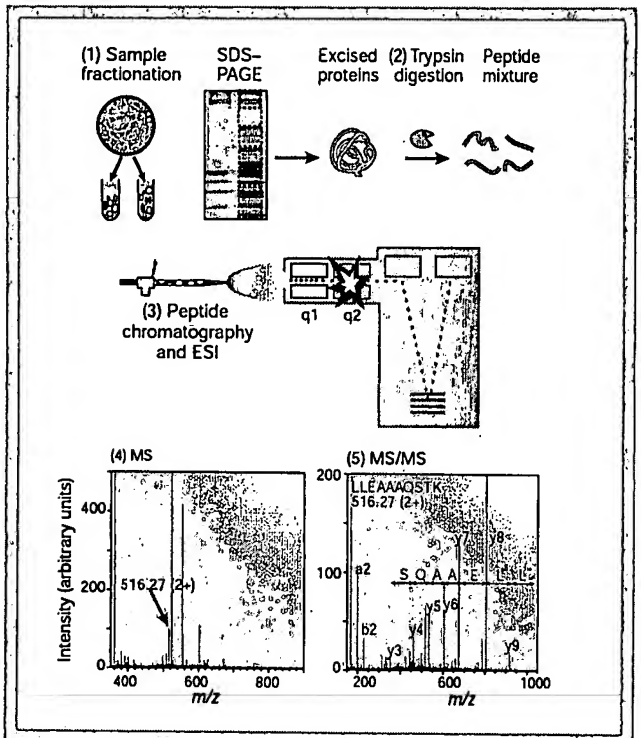


**Figure 1** Generic mass spectrometry (MS)-based proteomics experiment. The typical proteomics experiment consists of five stages. In stage 1, the proteins to be analysed are isolated from cell lysate or tissues by biochemical fractionation or affinity selection. This often includes a final step of one-dimensional gel electrophoresis, and defines the 'sub-proteome' to be analysed. MS of whole proteins is less sensitive than peptide MS and the mass of the intact protein by itself is insufficient for identification. Therefore, proteins are degraded enzymatically to peptides in stage 2, usually by trypsin, leading to peptides with C-terminally protonated amino acids, providing an advantage in subsequent peptide sequencing. In stage 3, the peptides are separated by one or more steps of high-pressure liquid chromatography in very fine capillaries and eluted into an electrospray ion source where they are nebulized in small, highly charged droplets. After evaporation, multiply protonated peptides enter the mass spectrometer and, in stage 4, a mass spectrum of the peptides eluting at this time point is taken (MS, 'spectrum, or 'normal mass spectrum'). The computer generates a prioritized list of these peptides for fragmentation and a series of tandem mass spectrometric or 'MS/MS' experiments ensues (stage 5). These consist of isolation of a given peptide ion, fragmentation by energetic collision with gas, and recording of the tandem or MS/MS spectrum. The MS and MS/MS spectra are typically acquired for about one second each and stored for matching against protein sequence databases. The outcome of the experiment is the identity of the peptides and therefore the proteins making up the purified protein population.

the traditional, three-dimensional ion traps, allowing increased sensitivity, resolution and mass accuracy. The FT-MS instrument is also a trapping mass spectrometer, although it captures the ions under high vacuum in a high magnetic field. Its strengths are high sensitivity, mass accuracy, resolution and dynamic range[8–11]. But in spite of the enormous potential, the expense, operational complexity and low peptide-fragmentation efficiency of FT-MS instruments has limited their routine use in proteomics research.

MALDI is usually coupled to TOF analysers that measure the mass of intact peptides, whereas ESI has mostly been coupled to ion traps and triple quadrupole instruments and used to generate fragment ion spectra (collision-induced (CID) spectra) of selected precursor ions[4]. More recently, new configurations of ion sources and mass analysers have found wide application for protein analysis. To allow the fragmentation of MALDI-generated precursor ions, MALDI ion sources have recently been coupled to quadrupole ion-trap mass spectrometers[12] and to two types of TOF instruments. In the first, two TOF sections are separated by a collision cell ('TOF-TOF instrument')[13], whereas in the second, the hybrid quadrupole TOF instrument, the collision cell is placed between a quadrupole mass filter and a TOF analyser[14]. Ions of a particular *m/z* are selected in a first mass analyser (TOF or quadrupole), fragmented in a collision cell and the fragment ion masses are 'read out' by a TOF analyser. These instruments have high sensitivity, resolution and mass accuracy, and the quadrupole TOF instrument can be used interchangeably with an ESI ionization source. The resulting fragment ion spectra are often more extensive and informative than those generated in trapping instruments. Although TOF, ion-trap and hybrid TOF instruments dominate proteomics today, other configurations including linear ion traps and FT-MS instruments could become widespread in the near future.

As a result of its simplicity, excellent mass accuracy, high resolution and sensitivity, MALDI-TOF is still much used to identify proteins by what is known as peptide mapping, also referred to as peptide-mass mapping or peptide-mass fingerprinting. In this method, proteins are identified by matching a list of experimental peptide masses with the calculated list of all peptide masses of each entry in a database (for example, a comprehensive protein database). Because mass mapping requires an essentially purified target protein, the technique is commonly used in conjunction with prior protein fractionation using either one- or two-dimensional gel electrophoresis (1DE and 2DE, respectively). The addition of sequencing capability to the MALDI method should make protein identifications by MALDI-MS/MS more specific than those obtained by simple peptide-mass mapping (see below). It should also extend the use of MALDI to the analysis of more complex samples, thereby uncoupling MALDI-MS from 2DE. However, if MALDI-MS/MS is to be used with peptide chromatography, the effluent of a liquid chromatography run must be deposited on a sample plate and mixed with the MALDI matrix, a process that has thus far proven difficult to automate. In general, it can be expected that the trend towards the combination of liquid chromatography with ESI- or MALDI-MS/MS (Fig. 1) will continue.

Protein identifications using peptide CID spectra are more clear-cut than those achieved by mass mapping because, in addition to the peptide mass, the peak pattern in the CID spectrum also provides information about peptide sequence. This information is not readily convertible into a full, unambiguous peptide sequence, that is, the '*de novo*' sequencing problem via MS is still not generally solved. Instead, the CID spectra are scanned against comprehensive protein sequence databases using one of a number of different algorithms, each with its strengths and weaknesses. The 'peptide sequence tag' approach extracts a short, unambiguous amino acid sequence from the peak pattern that, when combined with the mass information, is a specific probe to determine the origin of the peptide[15]. In the cross-correlation method, peptide sequences in the database are used to construct theoretical mass spectra and the overlap or 'cross-correlation' of these predicted spectra with the measured mass spectra determines the best match[16]. In the third main approach, 'probability based matching', the calculated fragments from peptide sequences in the database are compared with observed peaks. From this comparison a score is calculated which reflects the statistical significance of the match between the spectrum and the sequences contained in a database[17].

In each of these methods the identified peptides are compiled into a protein 'hit list', which is the output of a typical proteomic experiment. Because protein identifications rely on matches with sequence databases, high-throughput proteomics is currently restricted largely to those species for which comprehensive sequence databases are available.

## Protein identification and quantification

No method or instrument exists that is capable of identifying and quantifying the components of a complex protein sample in a simple, single-step operation. Rather, individual components for separating,

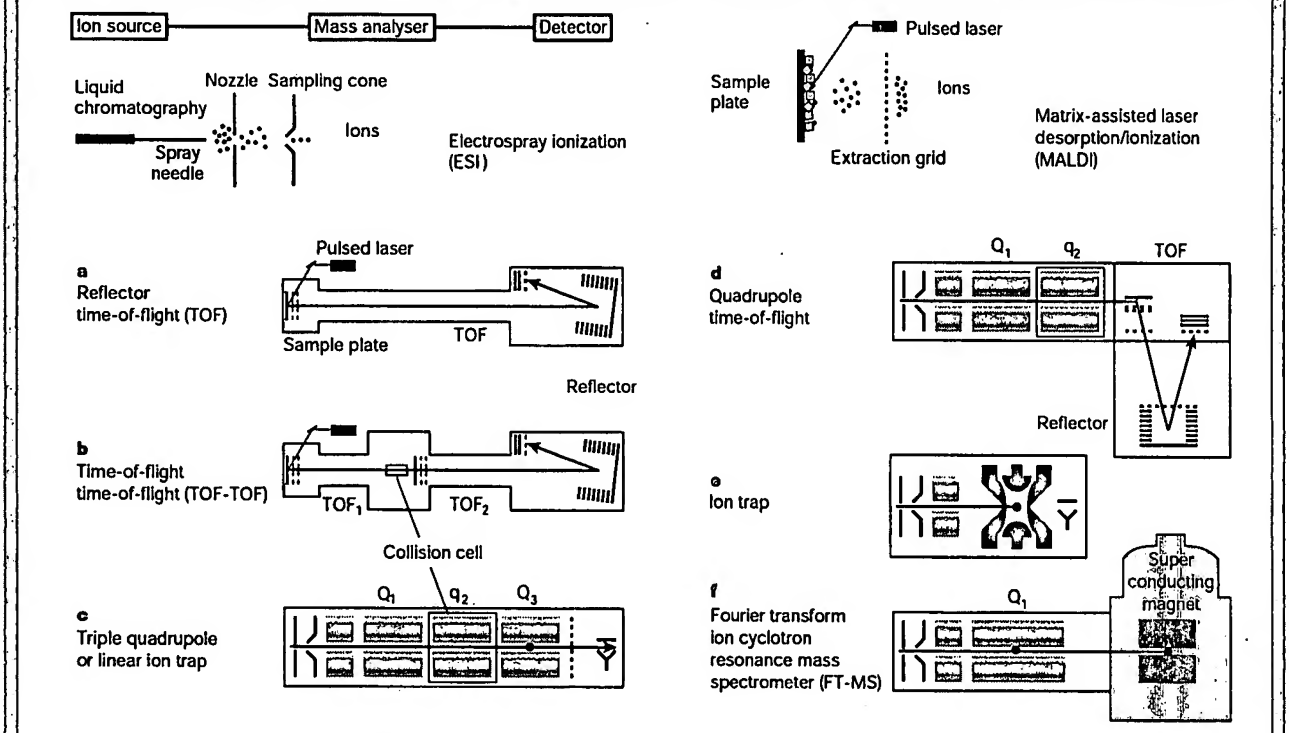


**Figure 2** Mass spectrometers used in proteome research. The left and right upper panels depict the ionization and sample introduction process in electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). The different instrumental configurations (**a–f**) are shown with their typical ion source. **a**, In reflector time-of-flight (TOF) instruments, the ions are accelerated to high kinetic energy and are separated along a flight tube as a result of their different velocities. The ions are turned around in a reflector, which compensates for slight differences in kinetic energy, and then impinge on a detector that amplifies and counts arriving ions. **b**, The TOF-TOF instrument incorporates a collision cell between two TOF sections. Ions of one mass-to-charge (*m/z*) ratio are selected in the first TOF section, fragmented in the collision cell, and the masses of the fragments are separated in the second TOF section. **c**, Quadrupole mass spectrometers select by time-varying electric fields between four rods, which permit a stable trajectory only for ions of a particular desired *m/z*. Again, ions of a particular m/z are selected in a first section ($Q_1$), fragmented in a collision cell ($q_2$), and the fragments separated in $Q_3$. In the linear ion trap, ions are captured in a quadruple section, depicted by the red dot in $Q_3$. They are then excited via resonant electric field and the fragments are scanned out, creating the tandem mass spectrum. **d**, The quadrupole TOF instrument combines the front part of a triple quadruple instrument with a reflector TOF section for measuring the mass of the ions. **e**, The (three-dimensional) ion trap captures the ions as in the case of the linear ion trap, fragments ions of a particular *m/z*, and then scans out the fragments to generate the tandem mass spectrum. **f**, The FT-MS instrument also traps the ions, but does so with the help of strong magnetic fields. The figure shows the combination of FT-MS with the linear ion trap for efficient isolation, fragmentation and fragment detection in the FT-MS section.

identifying and quantifying the polypeptides as well as tools for integrating and analysing all the data must be used in concert. Out of a bewildering multitude of techniques and instruments, two main tracks can be identified. The first, and most commonly used, is a combination of 2DE and MS. The second track combines limited protein purification with the more recently developed techniques of automated peptide MS/MS and, if accurate quantification is desired, stable-isotope tagging of proteins or peptides. In either track a suitable data processing, storage and visualization infrastructure needs to be developed, if the platform is intended for high-throughput operation.

In the first track, the proteins in a sample are separated by 2DE, stained, and each observed protein spot is quantified by its staining intensity. Selected spots are excised, digested and analysed by MS. Sophisticated pattern-matching algorithms as well as interpretation by skilled researchers are required to relate the 2DE patterns to each other in order to detect characteristic patterns and differences among samples. 2DE has been a mature technique for more than 25 years and was the first technique capable of supporting the concurrent quantitative analysis of large numbers of gene products. In fact, many of the principles now commonly used for global, quantitative analysis of messenger RNA expression patterns, such as clustering algorithms and multivariate statistics, were developed in the context of 2DE[18].

Peptide-mass mapping by MALDI-TOF and peptide sequencing by ESI-MS/MS have become highly efficient at the identification of gel-separated proteins. In the many reports using this technology, largely the same proteins were identified repeatedly, irrespective of the system studied, which suggests limited dynamic range of 2DE-based proteomics. Systematic studies of the budding yeast *Saccharomyces cerevisiae* indeed revealed that typically only the most abundant proteins can be observed by this method[19]. Incremental improvements in 2DE technology, including more sensitive staining methods[20,21], large-format higher resolving gels[22] and sample fractionation prior to 2DE have alleviated, but not eliminated, these and other shortcomings of the 2DE/MS approach.

Studying major histocompatibility complex class I-associated peptides, a natural and complex peptide library, Hunt and colleagues pioneered the use of LC-MS/MS for the analysis of complex peptide mixtures and it is this method that is today at the core of MS-based proteomics[23]. However, before LC-MS/MS could be used both for the identification of protein mixtures and for quantitative proteomic experiments, a number of technical issues had to be addressed. First, single-dimension peptide chromatography does not provide sufficient peak capacity to separate peptide mixtures as complex as those generated by the proteolysis of protein mixtures of, for example, total cell lysates. Second, in both MALDI- and ESI-MS, the relationship between the amount of analyte present and measured signal intensity is complex and incompletely understood. Mass spectrometers are therefore inherently poor quantitative devices.

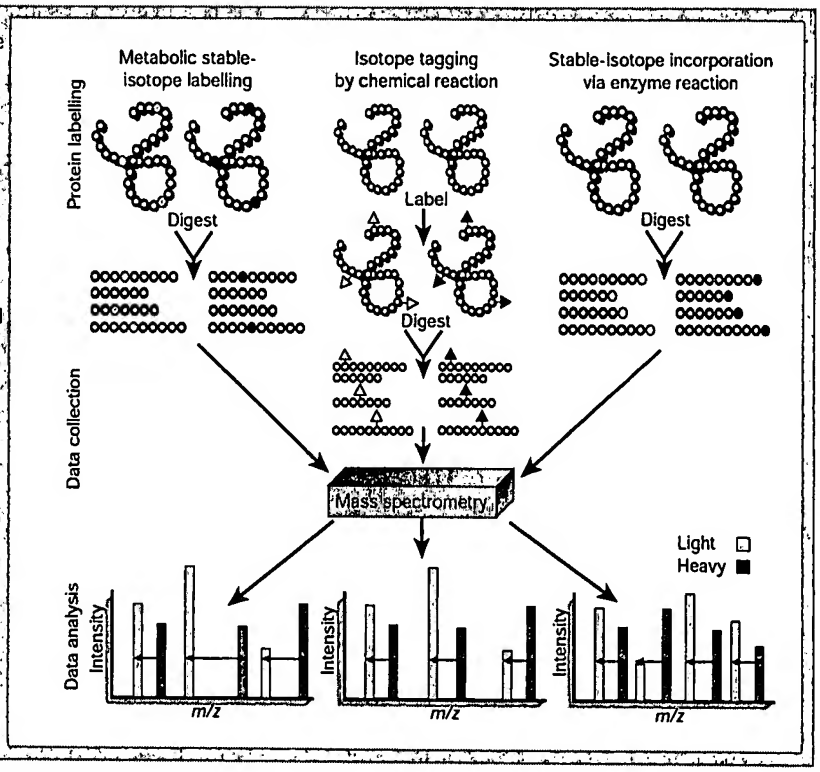


**Figure 3** Schematic representation of methods for stable-isotope protein labelling for quantitative proteomics. **a**, Proteins are labelled metabolically by culturing cells in media that are isotopically enriched (for example, containing $^{15}N$ salts, or $^{13}C$-labelled amino acids) or isotopically depleted. **b**, Proteins are labelled at specific sites with isotopically encoded reagents. The reagents can also contain affinity tags, allowing for the selective isolation of the labelled peptides after protein digestion. The use of chemistries of different specificity enables selective tagging of classes of proteins containing specific functional groups. **c**, Proteins are isotopically tagged by means of enzyme-catalysed incorporation of $^{18}O$ from $^{18}O$ water during proteolysis. Each peptide generated by the enzymatic reaction carried out in heavy water is labelled at the carboxy terminal. In each case, labelled proteins or peptides are combined, separated and analysed by mass spectrometry and/or tandem mass spectrometry for the purpose of identifying the proteins contained in the sample and determining their relative abundance. The patterns of isotopic mass differences generated by each method are indicated schematically. The mass difference of peptide pairs generated by metabolic labelling is dependent on the amino acid composition of the peptide and is therefore variable. The mass difference generated by enzymatic $^{18}O$ incorporation is either 4 Da or 2 Da, making quantitation difficult. The mass difference generated by chemical tagging is one or multiple times the mass difference encoded in the reagent used.

Third, the amount of data collected by the method is huge and its analysis daunting.

Substantial progress has been achieved in each of these areas, resulting in the emergence of increasingly robust and productive platforms. To provide more peak capacity, various combinations of protein and peptide separation schemes have been explored. Most popular at present are two-dimensional (strong cation exchange/reversed phase)[24,25] or three-dimensional (strong cation exchange/avidin/reversed phase)[28] chromatographic separations of peptide mixtures generated by tryptic digestion of protein samples that are frequently pre-fractionated by 1DE. Several studies suggest that, in principle, these methods are capable of detecting proteins of very low abundance, although considerable effort is required and a sufficient amount of starting protein sample must be available[27,28]. However, no proteome has yet been completely analysed and, for lack of a suitable reference, it will be difficult to determine when that milestone has been achieved.

To add a quantitative dimension to peptide LC-MS/MS experiments, the proven technique of stable-isotope dilution has been applied. This method makes use of the facts that pairs of chemically identical analytes of different stable-isotope composition can be differentiated in a mass spectrometer owing to their mass difference, and that the ratio of signal intensities for such analyte pairs accurately indicates the abundance ratio for the two analytes (Fig. 3). To this end, stable-isotope tags have been introduced to proteins via metabolic labelling using heavy salts or amino acids[29], enzymatically via transfer of $^{18}O$ from water to peptides[30,31], or via chemical reactions using isotope-coded affinity tags or similar reagents[32,33]. Post-isolation chemical isotope tagging of proteins is currently the most versatile and most commonly used labelling method. An attractive feature of this approach is that the selectivity of the labelling reactions can be used to direct the isotopes and attached affinity tags to specific functional groups or protein classes, thus enabling their selective isolation and analysis.

So far, isotope-tagging chemistries have been described that are specific for sulphydryl groups[32,33], amino groups[34], the active sites for serine[35] and cysteine hydrolases[36], for phosphate ester groups[37,38] and for *N*-linked carbohydrates[39]. Site-specific isotope tagging is limited only by the creativity of the chemist synthesizing suitable reagents. We therefore expect that new reagents will make many different types of 'sub-proteomes' accessible to quantitative analysis. Recently, a method called 'stable-isotope labelling with amino acids in cell culture', or SILAC, has been described[40]. In this method, one cell state is metabolically labelled by, for example, $^{13}C$-labelled arginine. Potentially all peptides can be labelled and the absence of any chemical steps make the method easy to apply as well as compatible with multistage purification procedures.

A current challenge for high-throughput proteomics is to use CID database search results from large numbers of peptide CID spectra to derive a list of identified peptides and their corresponding proteins. This task entails distinguishing correct peptide assignments from false identifications among database search results. In the case of small data sets, this can be achieved by researchers with expertise in spectral interpretation, manually verifying the peptide assignments to spectra made by database search programs. Such a time-consuming approach is not feasible for high-throughput analysis of large data sets containing tens of thousands of spectra, or when expertise is not available.

Alternatively, researchers can attempt to separate the correct from incorrect peptide assignments by applying filtering criteria based upon database search scores and other available data[25,26,28]. However, the rates of false identifications that result from such filters are not known, nor is it known how those rates are affected by mass spectrometer, sample preparation, or spectrum quality. In addition, researchers often use their own preferred filtering criteria, making it particularly difficult to compare their results among or even within groups. Consequently, the question of what constitutes an identified protein in a LC-MS/MS experiment has been difficult to answer. It is therefore important that computer programs that use robust and transparent statistical principles to estimate accurate probabilities indicating the likelihood for the presence of a peptide or protein in the sample[41,42] are further developed and widely tested and applied.

The technologies and tools described here are now being combined to create robust platforms for quantitative, high-throughput proteomics. This effort is aided by the introduction of the new types of high-performance mass spectrometers discussed

above. Currently, specialized MS laboratories can easily identify and quantify hundreds of proteins per day on a single MS system, and rapid advances in sample throughput, sensitivity and accuracy are projected.

## Applying proteomics technology to protein profiling

Protein mixtures of considerable complexity can now be routinely characterized in some depth using the methods described above. One measure of technical progress is the number of proteins identified in each study. Such numbers can now reach into the thousands for suitably complex samples. But to be biologically useful, as opposed to simply highlighting analytical features of the methods, large-scale proteomic studies need to solve biological questions. In this regard, MS-based proteomics has interfaced particularly well with three types of biological or clinical questions. The first is the generation of protein–protein linkage maps. The second is the use of protein identification technology to annotate and, if necessary, correct genomic DNA sequences. The third is the use of quantitative methods to analyse protein expression profiles as a function of cellular state as an aid to infer cellular function.

The sequences of many mature proteins in higher eukaryotes, after processing and splicing, are often not directly apparent from their cognate DNA sequences. Peptide sequence data of sufficient quality provides unambiguous evidence of translation of a particular gene and can, in principle, differentiate between alternatively spliced or translated forms of a protein. Using a combination of MS and gene chip analysis, a number of proteins that were derived from previously undetected open-reading frames were found in the yeast genome[43] and previously unknown human genes have also been found by direct searching of the human genome sequence[44,45].

Thus, it might be tempting to systematically analyse the proteins expressed by a cell or tissue, that is, to generate comprehensive proteome maps. First-generation large-scale proteome maps of microorganisms such as yeast[28] or the bacterium *Deinococcus radiodurans*[11] are examples of such projects and, with products from more than 60% of the genes identified, the *Deinococcus* map is at present the most complete. A recent review[46] of the proteomics of human plasma highlights a number of challenges facing comprehensive blood-serum analysis and thus, by implication, other samples from higher eukaryotes. Considering the combinatorial effects of splicing, processing and PTMs, plasma is estimated to contain many thousands to perhaps millions of polypeptide species, spanning a concentration range of up to 10 orders of magnitude. The fact that only about 500 proteins have so far been reported[47], and very few have been quantified, illustrates the need for further technological developments to address these issues.

The more common and versatile use of large-scale MS-based proteomics has been to document the expression of proteins as a function of cell or tissue state. We argue that to be meaningful, such data must be at least semi-quantitative and that a simple list of proteins detected in the different states is insufficient. This is because analyses of complex mixtures are often not comprehensive, and therefore the non-appearance of a particular sequence in the list of identified peptides does not indicate that the peptide or protein was not originally present in the sample. Additionally, it is often impossible to prepare a certain cell type, cell fraction or tissue in completely pure form, without trace contaminations of other fractions. And because the ion current of a peptide is dependent on a multitude of variables that are difficult to control, this measure is not a good indicator of peptide abundance. If stable-isotope dilution has not been used, a rough relative estimate of the quantity of the protein can be gained by integrating the ion current of its peptide-mass peaks over their elution time and comparing these 'extracted ion currents' between states, provided that highly accurate and reproducible methods are used.

The malaria parasite *Plasmodium falciparum* has recently been subjected to detailed proteomic analysis. The life cycle of the parasite is complex; thus there is great interest in the proteins it expresses in its own different stages and in its different host compartments. Illustrating the power and importance of proteomics, the recent genome project of the malaria parasite was accompanied by two large-scale proteome efforts. In one of the studies[48] the human stages of the parasite were analysed and a large number of proteins identified in the sexual and non-sexual stages. Quantitation was attempted by comparing peptide ion currents between stages and by correlation of protein data with RNA quantification by the polymerase chain reaction. After bioinformatic analysis, a set of proteins was selected from membrane fractions for follow-up as possible stage-specific drug or vaccine targets. The study resulted in more than 200 such candidate proteins and generated a large set of 'orphan' peptides that were not found in the set of predicted proteins, but were mapped onto the genome, assisting its annotation.

The other *P. falciparum* proteomics project analysed mosquito and human stages of the parasite and reported a total of around 2,400 identified proteins[49]. The study revealed unexpected stage specificity of a number of surface proteins and suggested co-expression of new proteins with groups of proteins already annotated as stage specific, helping to place these proteins in a functional context. The study also illustrated the need for transparent statistical tools to improve the confidence in protein identifications, as a large number, and in some cases the majority, of the proteins were identified solely by single peptides, many of which did not conform to the expected tryptic cleavage pattern.

Increasingly, stable-isotope dilution and LC-MS/MS are used to accurately detect changes in quantitative protein profiles and to infer biological function from the observed patterns. Shiio *et al.*[50] identified the reduction of stress fibres and focal adhesions as a new cellular function of the Myc oncogene by comparing protein extracts from $Myc^+$ and $Myc^-$ cells. Han *et al.*[26] identified pleiotropic, differentiation-induced effects in the microsomal compartment of phorbol ester-treated HL-60 cells, and a number of studies have also identified previously unknown connections between metabolic processes[51,52].

## Applying proteomics technology to protein interactions

The analysis of protein complexes is the third area where MS-based proteomics has had a significant impact. Most proteins exert their function by way of protein–protein interactions and enzymes are often held in tightly controlled regions of the cell by such interactions. Thus, one of the first questions usually asked about a new protein — apart from where it is expressed — is to what proteins does it bind? To study this question by MS, the protein itself is used as an affinity reagent to isolate its binding partners. Compared with two-hybrid and chip-based approaches, this strategy has the advantages that the fully processed and modified protein can serve as the bait, that the interactions take place in the native environment and cellular location, and that multicomponent complexes can be isolated and analysed in a single operation[53]. However, because many biologically relevant interactions are of low affinity, transient and generally dependent on the specific cellular environment in which they occur, MS-based methods in a straightforward affinity experiment will detect only a subset of the protein interactions that actually occur. Bioinformatics methods, correlation of MS data with those obtained by other methods, or iterative MS measurements possibly in conjunction with chemical crosslinking[54] can often help to further elucidate direct interactions and overall topology of multiprotein complexes.

MS-based protein interaction experiments have three essential components: bait presentation, affinity purification of the complex, and analysis of the bound proteins. Ideally, endogenous proteins can serve as bait if an antibody or other reagent exists that allows specific isolation of the protein with its bound partners. Unfortunately, there are currently no comprehensive antibody collections and many current antibodies do not immunoprecipitate well or lack sufficient specificity. A more generic strategy is to 'tag' the proteins of interest with a sequence readily recognized by an antibody specific for the tag.

To facilitate expression of the tagged protein at close to physiological levels, the tagged construct is preferably expressed from the promoter of its native, untagged counterpart. This can be achieved in a limited number of species, most notably *S. cerevisiae*, by using homologous recombination to replace the endogenous gene in the genome with a gene coding the tagged protein.

In mammalian cells, where expression of tagged proteins from the native promoter is more difficult, they are usually expressed after transient transfection, in stable cell lines generated by traditional selection, or by recently introduced kits for fast generation of stable cell lines. Transient or stable transfections usually result in tagged-protein expression levels that are different from the untagged, endogenous counterpart. They are therefore prone to artefacts generated by non-physiological levels of the bait protein. Considerable efforts have been devoted to developing tagging systems optimized for analysis of protein complexes (see review in this issue by Fields and co-workers, page 208). Tags supporting single-step purification have the advantage of convenience and yield. Tags supporting two sequential affinity steps (tandem affinity purification or TAP) combine two different tags on the same protein, which are normally separated by an enzyme-cleavable linker sequence.

A popular implementation of this concept consists of a calmodulin-binding domain in series with the immunoglobulin-binding domain of protein A, the domains being separated by a sequence that can be cleaved by a tobacco etch virus (TEV) protease[55]. The tagged proteins are bound initially to a solid support modified with immunoglobulins, recovered by TEV proteolysis and bound to a calmodulin column from which they can be selectively eluted by increased [$Ca^{2+}$]. TAP tags significantly reduce background noise, but probably result in the loss of some of the more transient and weak binding partners during the purification procedure, as the second affinity step essentially causes infinite sample dilution. The identification part of the strategy is similar to the generic protein identification experiment described above, and essentially all the strategies discussed here have been used for the analysis of protein complexes. However, it is clear that the cleaner the initial purification, the less challenging the mass spectrometric 'readout' becomes.

Combining such developments, two large-scale projects have recently been reported on the protein–protein interaction network in yeast. In one of the studies, 1,739 TAP-tagged genes were introduced into the yeast genome by homologous recombination, 232 stable complexes were isolated and protein constituents were identified by MALDI peptide mapping after separation by 1DE[56]. Apart from the large number of new interactions for known and new yeast proteins, a higher-order interaction structure between complexes emerged from the data. A similar study used transient transfection to express FLAG-tagged bait proteins; complexes were isolated by single-step immunopurification and any attached proteins identified by automated LC-MS/MS of gel-separated bands[57]. These experiments probed the phosphatases, kinases and the DNA-repair network of yeast specifically, resulting in many interesting signalling connections being made.

Both studies reported a large number of interacting proteins, and while groups of selected bait proteins were only partly overlapping, a number of interesting conclusions could be drawn from a comparison of the results. First, protein complexes isolated in a single step resulted in more complex samples than those isolated by the TAP tag procedure. Second, surprisingly little overlap of the data was observed when results from similar bait proteins were compared[58] between the two studies or between the MS and previous yeast two-hybrid studies. Although a variety of technical explanations have been advanced to explain this discrepancy, it is important to note that the 'interactome' is potentially very large, growing with the square of the number of proteins involved, and that it remains substantially undersampled. Third, both projects reported results consistent with previous literature for already known complexes. As in other large-scale projects, higher accuracy can be obtained with more detailed

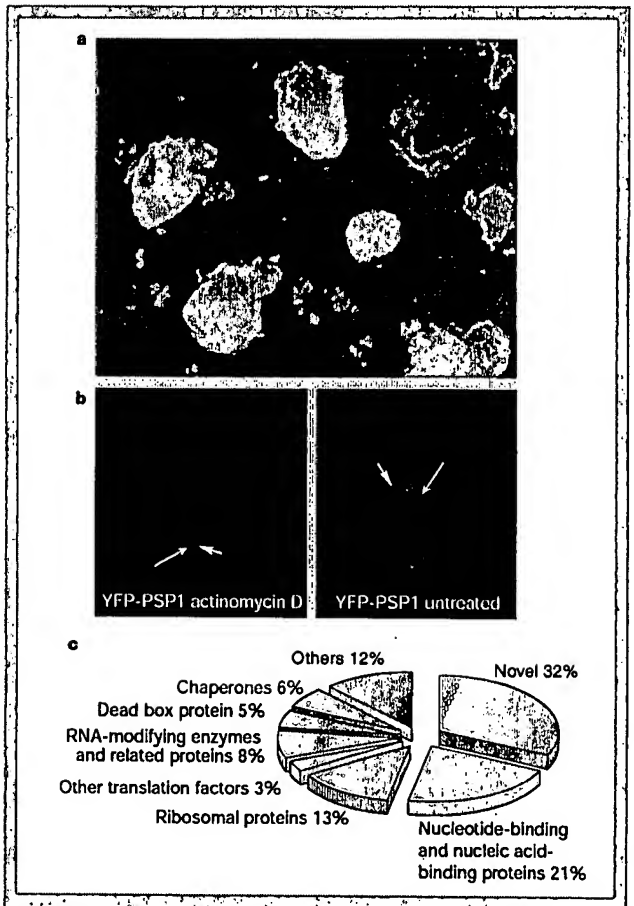


**Figure 4** Organellar proteomics by combined mass spectrometry (MS) and imaging methods. An organelle is isolated by biochemical methods such as density-gradient centrifugation. **a**, Scanning electron microscopy shows the result of purification of the human nucleolus. The organellar proteins are then solubilized, optionally fractionated, and analysed by the methods described in Fig. 1; this results in a list of proteins that are candidate constituents of the complex. At this stage MS-based proteomics is combined with imaging techniques such as laser confocal scanning fluorescence microscopy of yellow fluorescent protein (YFP)-tagged members of the complex. **b**, The right panel shows localization of one of the new proteins found by proteomics to subnuclear structures termed 'paraspeckles'. The left panel shows that inhibition of transcription by actinomycin D leads to reorganization of the nucleoli and concentration of the novel protein to cap structures[91]. **c**, Functional categorization of the proteins in the organelle. More than 400 proteins have been found and can be assayed for dynamic localization and biochemical function in the context of the nucleolus.

experiments, for example, in a 'complex walking' strategy in which complexes are tagged and identified sequentially[59].

In the future, quantitative methods based on stable-isotope labelling are likely to revolutionize the study of stable or transient interactions and interactions dependent on PTMs. In such experiments, accurate quantification by means of stable-isotope labelling is not used for protein quantification per se; instead the stable-isotope ratios distinguish between the protein composition of two or more protein complexes. In the case of a sample containing a complex and a control sample containing only contaminating proteins (for example, immunoprecipitation with an irrelevant antibody or isolate from a cell devoid of affinity-tagged protein), the method can distinguish between true complex components and nonspecifically associated proteins. In the case of complexes isolated from cells at different states (for example, activated and non-activated cells) the method can identify dynamic changes in the composition of protein complexes[60,61].

The ability of quantitative MS to detect specific complex components within a background of nonspecifically associated proteins increases the tolerance for high background and allows for fewer purification steps and less stringent washing conditions, thus increasing the chance of finding transient and weak interactions. The same methods can be used to study the interaction of proteins with nucleic acids, small molecules and in fact with any other substrate. For example, drugs can be used as affinity baits in the same way as proteins to define their cellular targets, and small molecules such as co-factors can be used to isolate interesting 'sub-proteomes'[62].

MS-based proteomics is not limited to the analysis of complexes consisting of only a few proteins. In fact, some of the most biologically informative results have come from the analysis of large protein complexes — 'molecular machines', organelles and subcompartments of the cell. The first complex analysed in this way was the spliceosome, studied in yeast[63] and then in human cells[64], closely followed by the yeast nuclear pore complex[65]. Re-analysis of the spliceosome using more complete databases and more advanced instrumentation has recently been undertaken. In one study, nearly 300 proteins were found, and evidence from sequence analysis highlighted a set of 55 novel proteins involved in splicing and RNA processing[66]. A similar study, using an elegant RNA tag-based purification, also discovered many new proteins[67]. Both studies found essentially the complete list of known human splicing factors. The new data encompassed and extended the original results, indicating the maturity of MS-based methods for the analysis of such complex structures. The next challenge will now be to study the dynamics and assembly of functional protein modules via quantitative proteomics.

Numerous other large complexes and organelles have now at least been partly characterized by MS[68]. The limiting factor in such experiments is no longer primarily the analysis, but rather the ability to purify such structures to homogeneity. For example, it is very difficult to isolate structures such as the Golgi apparatus and the interpretation of results from samples of dubious quality and definition is correspondingly vague. The largest organelle mapped so far is the human nucleolus, whose high specific density allows for a simple, efficient purification[45] (Fig. 4). By using a variety of mass spectrometric techniques, more than 400 nucleolar proteins have now been identified in this structure. Well-characterized proteins identified in this study, but not previously known to be associated with nucleolus, raise interesting questions about the function of this organelle, while the identification of a large number of previously uncharacterized gene products places many of those in the context of nucleolar function.

At the same time, some of the previously known nucleolar proteins, such as Werner's syndrome protein, have not yet been found, indicating that even this large-scale study is not yet complete. One reason for this is that numerous factors, including Werner's syndrome protein, exhibit either cell-cycle dependent or facultative interactions with nucleoli. Dynamic imaging studies of the nucleus also make it clear that many of the factors in the nucleolus are associated only transiently with this organelle[69], a fact reflected in the overlapping 'cast of characters' of several nuclear bodies studied. Just as the single protein/single function concept is turning out to be more the exception than the rule, the concept of a single subcellular location of a protein may also turn out to be a gross over-simplification.

## Applying proteomics to the analysis of protein modifications

Proteins are converted to their mature form through a complicated sequence of post-translational protein processing and 'decoration' events. Many of the PTMs are regulatory and reversible, most notably protein phosphorylation, which controls biological function through a multitude of mechanisms. Mass spectrometric methods to determine the type and site of such modifications on single, purified proteins have been refined over the past two decades. In this case, peptide mapping with different enzymes is usually used to 'cover' as much of the protein sequence as possible. Protein modifications are then determined by examining the measured mass and fragmentation spectra via manual or computer-assisted interpretation. For the analysis of some types of PTMs, specific mass spectrometric techniques have been developed that scan the peptides derived from a protein for the presence of a particular modification. The analysis of regulatory modifications, in particular protein phosphorylation, is complicated by the frequently low stoichiometry, the size and ionizability of peptides bearing the modifications, and their fragmentation behaviour in the mass spectrometer (reviewed in refs 4,70,71). The analysis of the modification state of a purified protein therefore remains a challenging analytical endeavour.

Recently, attempts have been made to define modifications on a proteome-wide scale. Given the difficulties of identifying all modifications even in a single protein, it is clear that, at present, scanning for proteome-wide modifications is not comprehensive. Nevertheless, a large amount of biologically useful information can, in principle, be generated by this approach. One of the strategies used is essentially an extension of the approach used to analyse protein mixtures[72]. Instead of searching the database only for non-modified peptides, the database search algorithm is instructed to also match potentially modified peptides. To avoid a 'combinatorial explosion' resulting from the need to consider all possible modifications for all peptides in the database, the experiment is usually divided into identification of a set of proteins on the basis of non-modified peptides, followed by searching only these proteins for modified peptides[72].

A more functionally oriented strategy focuses on the search for one type of modification on all the proteins present in a sample. Such techniques are based usually on some form of affinity selection that is specific for the modification of interest and which is used to purify the 'sub-proteome' bearing this modification. For example, Pandey *et al.* stimulated cells with epidermal growth factor and isolated newly phosphotyrosine-modified proteins using antibodies specific to phosphotyrosine[73,74]. Efforts to determine the 'phosphoproteome' in a single step[37,38] have used chemical modification combined with affinity selection. In a potentially powerful technique, Ficcaro *et al.*[75] esterified peptide mixtures, thereby nullifying negatively charged carboxyl groups, and then captured phosphopeptides on metal affinity columns. This approach overcomes the low specificity of these columns caused by their affinity for any negatively charged peptides and seems to significantly improve the capture of phosphopeptides. Further development of these and related techniques may allow study of the complete phosphoproteome in multiprotein complexes and the pattern of the more abundant phosphopeptides in whole cells, a promising approach to study the activation state of whole signalling networks.

Gygi *et al.* have used affinity purification to capture the ubiquitinated proteins of yeast cells (ref. 76 and S. P. Gygi, personal communication). Over 1,000 such proteins were identified and in more than 100 cases the site of ubiquitination was determined. These results open up the study of ubiquitinated substrates in a cell state- and protein complex-dependent manner.

Many challenges remain in the large-scale mapping of PTMs, but it is clear that MS-based proteomics can make a unique contribution in this area. For example, systematic quantitative measurements of PTMs by stable-isotope labelling would be of tremendous biological interest.

## Challenges, expectations and emerging technologies

Proteomics, in particular quantitative proteomics, can be viewed as an array of biological or clinical assays capable of probing most, if not all, of the proteins in a sample. As proteins are involved in essentially all biological functions and clinical conditions, MS and proteomics will have an even greater impact on biology and medicine than it has had so far.

Over the past decade, MS of single proteins or protein complexes has been successful to the point where it is now considered a mainstream technology. This technology interfaces particularly well with biochemical and cell biological studies for studying specific protein functions. The success is built on the proven potential of mass spectrometric techniques to rapidly identify almost any protein, to

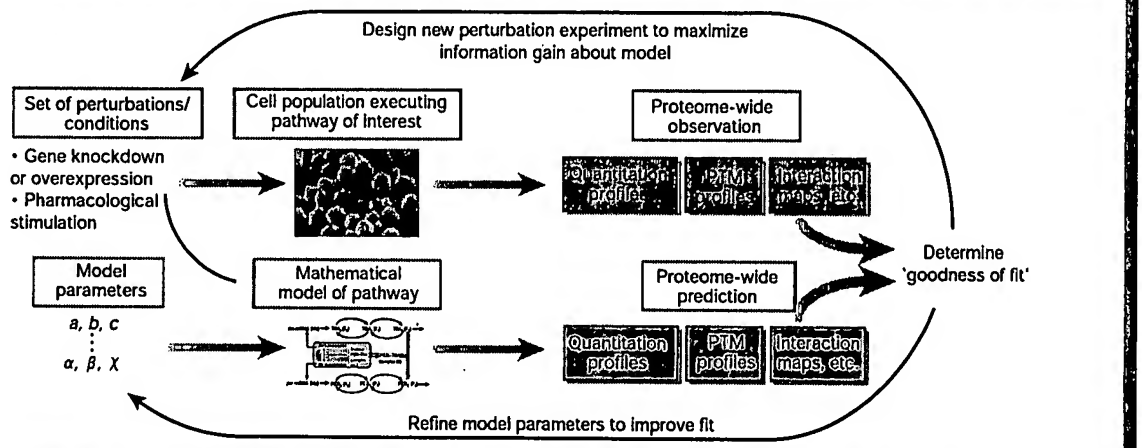


**Figure 5** Schematic representation of the systems biology paradigm. Cells are subjected to specific (for example, genetic or pharmacological) perturbations within the space of the system studied and the effects of the perturbations on the cells are recorded using systematic genomic and proteomic methods of analysis. Proteomic data that are particularly informative include quantitative protein profiles, profiles of regulatory modifications and protein interaction networks. The data are integrated and reconciled with prior models describing the system studied and discrepancies between the observed data and the model are used to design new perturbations, which are analysed by means of systematic measurements. The process is repeated iteratively until model and observed data converge.

analyse that protein for the presence of PTMs, to determine how and with what other biomolecules that proteins interacts, and even to gain structural information about the protein from gas-phase experiments[77,78] and from experiments in which mass spectrometric characterization of proteins has interfaced with X-ray crystallography[79]. As analytical methods and instrumentation are improving constantly, MS can be use to address an increasing number of analytical problems facing biochemists, geneticists and cell biologists. But protein MS does not equal proteomics. The specific objective of proteomics is to concurrently identify, quantify and analyse a large number of proteins in a functional context. This shift in focus from the analysis of selected isolated proteins to proteome-wide analyses has a number of profound implications and poses as yet unmet challenges for every aspect of experimental biology. These include experimental design, data analysis, visualization and storage, organization of proteomics research groups and publication of proteomic data.

### Experimental design

In a typical protein MS experiment, a specific property (for example, sequence, PTM or interaction) of a partly characterized protein is examined. In contrast, proteomic experiments often collect large amounts of data in the absence of hypotheses concerning specific proteins or activities. Proteomic experiments, therefore, have to be designed in ways that maximize the likelihood of generating new discoveries, or at least new testable hypotheses. The technology of gene expression profiling is conceptually similar to proteomic profiling and has demonstrated that more information is better. Although it is essentially impossible to draw meaningful conclusions from a single quantitative gene expression profile, the availability of multiple profiles from related samples allows the application of statistical tools[80] to extract signature patterns containing diagnostic or functional information. Therefore, successful proteomics experiments need to be designed in such a way that they can take advantage of the power of statistics for data interpretation. To achieve this goal, carefully controlled repeat studies and the generation of models describing the source, magnitude and distribution of errors will be essential.

### Data collection

Proteomic studies necessarily result in large amounts of data. Data collection at a volume and quality that is consistent with the use of statistical methods is a significant limitation of proteomics today. In a typical LC-MS/MS experiment, approximately 1,000 CID spectra can be acquired per hour. Even with the optimistic assumption that every one of these spectra leads to the successful identification of a peptide, it would take a long time to analyse complete proteomes. High-throughput collection of consistently high-quality data therefore remains a challenge in proteomics. We have argued that one solution to the problem would be to establish a number of specialized and generally accessible data-collection centres[81], akin to the beam lines used by X-ray crystallographers for protein structural studies. Such centres would not only generate data of consistent quality for a large number of proteomics projects, but would also serve as disseminators of advanced technology.

### Data analysis, visualization and storage

The analysis and interpretation of the enormous volumes of proteomic data remains an unsolved challenge, particularly for gel-free approaches. Expert manual analysis is incompatible with the tens of thousands of spectra collected in a single experiment and is inconsistent. Therefore, the development of transparent tools for the analysis of proteomic data using statistical principles is a key challenge[41,42]. Only once such tools are tested, validated and widely accepted will it become feasible to apply quality standards for protein identification, quantification and other measurements and to compare complementary proteomic data sets generated in different laboratories. These comparisons will also depend critically on transparent file structures for data storage, communication and visualization. The development of such proteomics tools is still in its infancy.

### Data publication

The publication of the large data sets generated by proteomic experiments and the information contained therein poses significant challenges. At present, most proteomics publications consist of a experimental description, a data table (typically published as supplementary material containing a partially interpreted and validated summary of the data) and an in-depth validation and discussion of one to a few conclusions made from the data. To make publication of proteomics data more useful, publishers and journals need to find new ways to review large data sets, to validate their contents and to make the information contained therein electronically searchable;

this problem remains essentially unsolved, despite preliminary developments by a few publishers and journals[82].

In spite of these and other challenges, the impact of proteomics on clinical and biological research is growing rapidly. It seems that beyond its great current contribution to cell biology, proteomics may have a huge influence on clinical diagnosis. MS-based proteomics seems capable of detecting patterns of differentially expressed proteins in easily accessible clinical samples such as blood serum. These types of analyses have the potential to diagnose the presence and stage of many diseases, in particular cancers[83]. Clinical diagnosis will be further advanced with the advent of mass spectrometers with higher mass accuracy, dynamic range and resolution, and with the ability to identify specific sequences of diagnostic analytes and the use of accurate quantification procedures.

MS-based proteomics is still an emerging technology where revolutionary change is possible. Several concepts have been proposed and are under development that have the potential to alter the landscape of current MS-based proteomic technologies. One of these is the analysis of intact proteins. The currency of essentially all MS-based identifications is peptides. The convergence of mass spectrometers with large mass ranges, extremely high mass accuracy and resolution, and ionization/fragmentation methods compatible with large proteins has catalysed the emergence of whole-protein proteomics[84]. The analysis of whole proteins with high accuracy has the potential to distinguish and characterize differentially modified forms and to provide insights into coordinated modification patterns that are difficult to establish by peptide analysis.

A second emerging concept is mass spectrometric tissue imaging[85]. In this technique, thin tissue sections are directly applied to a MALDI mass spectrometer and, after treating the samples with a suitable matrix, profiles of the proteins contained in the section are generated by 'imaging' the sample with an array of mass spectra. The method, while currently incapable of identifying the detected protein features, has already provided proof-of-principle that clinically diagnostic patterns can be generated. Increased spatial resolution, potentially to subcellular levels, improved software tools and automated sample preparation will further increase the utility of this technique for clinical diagnosis and classification.

A third concept is the use of mass tags measured in mass spectrometers of very high mass accuracy and resolution such as FT-MS instruments. These mass tags could be used potentially for high-throughput protein identification. The idea is based on the observation that a particular proteome, if digested with a specific enzyme such as trypsin, will generate a peptide mixture in which most peptides can be uniquely classified based on their accurate mass and some other parameters such as chromatographic coordinates[86,87]. Therefore, once the peptides are identified by MS/MS and annotated with accurate mass tags they can be identified in subsequent experiments simply by correlating the accurate mass and the separation coordinates with the list of previously determined mass tags.

## Conclusion and perspective

In studying a biological system using the biochemical approach, researchers have traditionally attempted to purify to homogeneity each of the system's components; each element is then studied in detail with the ultimate aim being to reconstitute the system *in vitro* from the isolated components. Because proteins carry out most biological activities, the biochemical approach has been significantly enhanced by the availability of the sensitive and rapid MS-based protein identification methods discussed in this article. The availability of complete genomic sequences from a number of species further facilitates MS-based protein identifications, as the requirement for *de novo* sequencing has been usurped by simple correlation of measured data versus theoretical data predicted from sequence databases. The availability of completely sequenced genomes also catalysed the emergence of systems biology — the attempt to systematically study all the concurrent physiological processes in a cell or tissue by global measurement of differentially perturbed states (Fig. 5). The ultimate goal of systems biology is the integration of data from these observations into models that might, eventually, represent and simulate the physiology of the cell[88].

Proteomics is an essential component of systems biology research because proteins are rich in information that has turned out to be extremely valuable for the description of biological processes. These include protein abundances, linkage maps to other proteins or to other types of biomolecules including DNA and lipids, activities, modification states, subcellular location and more. Unfortunately, with the exception of quantitative protein profiles and protein–protein interactions (keeping in mind the caveats discussed above), none of these properties can currently be measured systematically, quantitatively and with high throughput. But rapid advances in technology suggest that this limitation may be transient. The few studies where the same biological system was subjected to different types of systematic measurements already offer insights into the power of the method. For instance, mRNA expression profiles and protein expression profiles seem to be largely complementary and therefore contribute to a more refined description of the system that each observation by itself is unable to provide[88].

Extrapolating from these limited studies, we expect that combining different genomic and proteomic results obtained from the same biological system will substantially increase our understanding of complex biological processes. More specifically, the systems biology studies based on diverse and high-quality proteomic data will define functional biological modules, reveal previously unrecognized connections between biochemical processes and modules, and generate new hypotheses that can be tested either by traditional methods or by the targeted generation of more genomic and proteomic data[51,88–90]. □

doi:10.1038/nature01511


1. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* 405, 837–846 (2000).
2. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for the mass spectrometry of large biomolecules. *Science* 246, 64–71 (1989).
3. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular mass exceeding 10000 daltons. *Anal. Chem.* 60, 2299–2301 (1988).
4. Aebersold, R. & Goodlett, D. R. Mass spectrometry in proteomics. *Chem. Rev.* 101, 269–295 (2001).
5. Mann, M., Hendrickson, R. C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* 70, 437–473 (2001).
6. Hager, J. W. A new linear ion trap mass spectrometer. *Rapid Commun. Mass Spectrom.* 16, 512–526 (2002).
7. Schwartz, J. C., Senko, M. W. & Syka, J. E. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* 13, 659–669 (2002).
8. Marshall, A. G., Hendrickson, C. L. & Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom. Rev.* 17, 1–35 (1998).
9. Valaskovic, G. A., Kelleher, N. L. & McLafferty, F. W. Attomole protein characterization by capillary electrophoresis-mass spectrometry. *Science* 273, 1199–2202 (1996).
10. Martin, S. E., Shabanowitz, J., Hunt, D. F. & Marto, J. A. Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* 72, 4266–4274 (2000).
11. Lipton, M. S. *et al.* Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl Acad. Sci. USA* 99, 11049–11054 (2002).
12. Krutchinsky, A. N., Kalkum, M. & Chait, B. T. Automatic identification of proteins with a MALDI-quadrupole ion trap mass spectrometer. *Anal. Chem.* 73, 5066–5077 (2001).
13. Medzihradszky, K. F. *et al.* The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal. Chem.* 72, 552–558 (2000).
14. Loboda, A. V., Krutchinsky, A. N., Bromirski, M., Ens, W. & Standing, K. G. A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser desorption/ionization source: design and performance. *Rapid Commun. Mass Spectrom.* 14, 1047–1057 (2000).
15. Mann, M. & Wilm, M. S. Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399 (1994).
16. Eng, J. K., McCormack, A. L. & Yates, J. R. I An approach to correlate MS/MS data to amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5, 976–989 (1994).
17. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567 (1999).
18. Anderson, N. L., Hofmann, J. P., Gemmell, A. & Taylor, J. Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clin. Chem.* 30, 2031–2036 (1984).
19. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl Acad. Sci. USA* 97, 9390–9395 (2000).
20. Rabilloud, T. Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* 2, 3–10 (2002).
21. Unlu, M., Morgan, M. E. & Minden, J. S. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* 18, 2071–2077 (1997).

22. Gauss, C., Kalkum, M., Lowe, M., Lehrach, H. & Klose, J. Analysis of the mouse proteome. (I) Brain proteins: separation by two-dimensional electrophoresis and identification by mass spectrometry and genetic variation. *Electrophoresis* **20**, 575–600 (1999).
23. Hunt, D. F. *et al.* Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263 (1992).
24. Wolters, D. A., Washburn, M. P. & Yates, J. R. III An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).
25. Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.* **17**, 676–682 (1999).
26. Han, D. K., Eng, J., Zhou, H. & Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnol.* **19**, 946–951 (2001).
27. Gygi, S. P., Rist, B., Griffin, T. J., Eng, J. & Aebersold, R. Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags. *J. Proteome Res.* **1**, 47–54 (2002).
28. Washburn, M. P., Wolters, D. & Yates, J. R. III Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19**, 242–247 (2001).
29. Conrads, T. P., Issaq, H. J. & Veenstra, T. D. New tools for quantitative phosphoproteome analysis. *Biochem. Biophys. Res. Commun.* **290**, 885–890 (2002).
30. Mirgorodskaya, O. A. *et al.* Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using $^{18}O$-labeled internal standards. *Rapid Commun. Mass Spectrom.* **14**, 1226–1232 (2000).
31. Yao, X., Freas, A., Ramirez, J., Demirev, P. A. & Fenselau, C. Proteolytic $^{18}O$ labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal. Chem.* **73**, 2836–2842 (2001).
32. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17**, 994–999 (1999).
33. Zhou, H., Ranish, J. A., Watts, J. D. & Aebersold, R. Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nature Biotechnol.* **20**, 512–515 (2002).
34. Munchbach, M., Quadroni, M., Miotto, G. & James, P. Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal. Chem.* **72**, 4047–4057 (2000).
35. Liu, Y., Patricelli, M. P. & Cravatt, B. F. Activity-based protein profiling: the serine hydrolases. *Proc. Natl Acad. Sci. USA* **96**, 14694–14699 (1999).
36. Greenbaum, D., Medzihradszky, K. F., Burlingame, A. & Bogyo, M. Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. *Chem. Biol.* **7**, 569–581 (2000).
37. Zhou, H., Watts, J. D. & Aebersold, R. A systematic approach to the analysis of protein phosphorylation. *Nature Biotechnol.* **19**, 375–378 (2001).
38. Oda, Y., Nagasu, T. & Chait, B. T. Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nature Biotechnol.* **19**, 379–382 (2001).
39. Zhang, H., Li, X.-J., Martin, D. & Aebersold, R. Quantitative analysis of glycoproteins: applications to serum and membrane proteins. (submitted).
40. Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
41. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
42. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* DOI: 10.1021/pr025556v (2002).
43. Oshiro, G. *et al.* Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res.* **12**, 1210–1220 (2002).
44. Kuster, B., Mortensen, P., Andersen, J. S. & Mann, M. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1**, 641–650 (2001).
45. Andersen, J. S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12**, 1–11 (2002).
46. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
47. Adkins, J. N. *et al.* Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol. Cell. Proteomics* DOI: 10.1074/mcp.M200066-MCP200 (2002).
48. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419**, 537–542 (2002).
49. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526 (2002).
50. Shilo, Y. *et al.* Quantitative proteomic analysis of Myc oncoprotein function. *EMBO J.* **21**, 5088–5096 (2002).
51. Griffin, T. J. *et al.* Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **1**, 323–333 (2002).
52. Baliga, N. S. *et al.* Coordinate regulation of energy transduction modules in *Halobacterium* sp. analyzed by a global systems approach. *Proc. Natl Acad. Sci. USA* **99**, 14913–14918 (2002).
53. Ashman, K., Moran, M. F., Sicheri, F., Pawson, T. & Tyers, M. Cell signalling—the proteomics of it all. *Science's STKE* <http://stke.sciencemag.org/cgi/content/full/sigtrans;2001/103/pe33> (2001).
54. Rappsilber, J., Siniossoglou, S., Hurt, E. C. & Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal. Chem.* **72**, 267–275 (2000).
55. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17**, 1030–1032 (1999).
56. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
57. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
58. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
59. Shevchenko, A., Schaft, D., Roguev, A., Pijnappel, W. W. & Stewart, A. F. Deciphering protein complexes and protein interaction networks by tandem affinity purification and mass spectrometry: analytical perspective. *Mol. Cell. Proteomics* **1**, 204–212 (2002).
60. Blagoev, B. *et al.* A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nature Biotechnol.* advance online publication, 10 February 2003 (doi:10.1038/nbt790).
61. Ranish, J. A. *et al.* The study of macromolecular complexes by quantitative proteomics. *Nature Genet.* (in the press).
62. MacDonald, J. A., Mackey, A. J., Pearson, W. R. & Haystead, T. A. A strategy for the rapid identification of phosphorylation sites in the phosphoproteome. *Mol. Cell. Proteomics* **1**, 314–322 (2002).
63. Neubauer, G. *et al.* Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl Acad. Sci. USA* **94**, 385–390 (1997).
64. Neubauer, G. *et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature Genet.* **20**, 46–50 (1998).
65. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148**, 635–651 (2000).
66. Rappsilber, J., Ryder, U., Lamond, A. I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12**, 1231–1245 (2002).
67. Zhou, Z., Licklider, L. J., Gygi, S. P. & Reed, R. Comprehensive proteomic analysis of the human spliceosome. *Nature* **419**, 182–185 (2002).
68. Taylor, S. W., Fahy, E. & Ghosh, S. S. Global organellar proteomics. *Trends Biotechnol.* **21**, 82–88 (2003).
69. Leung, A. K. & Lamond, A. I. In vivo analysis of NHPX reveals a novel nucleolar localization pathway involving a transient accumulation in splicing speckles. *J. Cell Biol.* **157**, 615–629 (2002).
70. Mann, M. *et al.* Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol.* **20**, 261–268 (2002).
71. Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nature Biotechnol.* (in the press).
72. MacCoss, M. J. *et al.* Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl Acad. Sci. USA* **99**, 7900–7905 (2002).
73. Pandey, A. *et al.* Analysis of receptor signaling pathways by mass spectrometry: identification of Vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proc. Natl Acad. Sci. USA* **97**, 179–184 (2000).
74. Steen, H., Kuster, B., Fernandez, M., Pandey, A. & Mann, M. Tyrosine phosphorylation mapping of the epidermal growth factor receptor signaling pathway. *J. Biol. Chem.* **277**, 1031–1039 (2002).
75. Ficarro, S. B. *et al.* Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nature Biotechnol.* **20**, 301–305 (2002).
76. Peng, J. & Gygi, S. P. Proteomics: the move to mixtures. *J. Mass Spectrom.* **36**, 1083–1091 (2001).
77. Hanson, C. L., Fucini, P., Ilag, L. L., Nierhaus, K. H. & Robinson, C. V. Dissociation of intact *Escherichia coli* ribosomes in a mass spectrometer—evidence for conformational change in a ribosome elongation factor G complex. *J. Biol. Chem.* **278**, 1259–1267 (2002).
78. Oh, H. *et al.* Secondary and tertiary structures of gaseous protein ions characterized by electron capture dissociation mass spectrometry and photofragment spectroscopy. *Proc. Natl Acad. Sci. USA* **99**, 15863–15868 (2002).
79. Cohen, S. L. & Chait, B. T. Mass spectrometry as a tool for protein crystallography. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 67–85 (2001).
80. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
81. Aebersold, R. & Watts, J. D. The need for national centers for proteomics. *Nature Biotechnol.* **20**, 651 (2002).
82. Mann, M. A home for proteomics data? *Nature* **420**, 21 (2002).
83. Petricoin, E. F. *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577 (2002).
84. Mørtz, E. *et al.* Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl Acad. Sci. USA* **93**, 8264–8267 (1996).
85. Stoeckli, M., Chaurand, P., Hallahan, D. E. & Caprioli, R. M. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nature Med.* **7**, 493–496 (2001).
86. Goodlett, D. R. *et al.* Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching. *Anal. Chem.* **73**, 1112–1118 (2000).
87. Smith, R. D. *et al.* An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2**, 513–523 (2002).
88. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
89. Betts, J. C., Lukey, P. T., Robb, L. C., McAdam, R. A. & Duncan, K. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Mol. Microbiol.* **43**, 717–731 (2002).
90. Guina, T. *et al.* Quantitative proteomic analysis of *Pseudomonas aeruginosa* indicates synthesis of quinolone signal in adaptation to cystic fibrosis airways. *Proc. Natl Acad. Sci. USA* (in the press).
91. Fox, A. H. *et al.* Paraspeckles. A novel nuclear domain. *Curr. Biol.* **12**, 13–25 (2002).


**Acknowledgements** We thank members of the Institute of Systems Biology (ISB) and the Center for Experimental BioInformatics (CEBI) for critical reading of the manuscript, preparation of figures and fruitful discussions, especially L. Foster, S.-E. Ong, J. Andersen and L. Feltz. CEBI is supported by a grant from the Danish Natural Research Foundation. R.A. is supported by grants from the National Institute of Health and Oxford GlycoSciences, and a contract from the National Heart, Lung, and Blood Institute, National Institutes of Health. The ISB is supported in part by a gift from Merck and Co.